Fast Structural Similarity Search of Noncoding RNAs Based on Matched Filtering of Stem Patterns

Byung-Jun Yoon
Dept. of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
Email: bjyoon@caltech.edu

P. P. Vaidyanathan
Dept. of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
Email: ppvnath@systems.caltech.edu

Abstract— **Many noncoding RNAs (ncRNAs) have characteristic secondary structures that give rise to complicated base correlations in their primary sequences. Therefore, when performing an RNA similarity search to find new members of a ncRNA family, we need a statistical model – such as the profile-csHMM or the covariance model (CM) – that can effectively describe the correlations between distant bases. However, these models are computationally expensive, making the resulting RNA search very slow. To overcome this problem, various prescreening methods have been proposed that first use a simpler model to scan the database and filter out the dissimilar regions. Only the remaining regions that bear some similarity are passed to a more complex model for closer inspection. It has been shown that the prescreening approach can make the search speed significantly faster at no (or a slight) loss of prediction accuracy. In this paper, we propose a novel prescreening method based on matched filtering of stem patterns. Unlike many existing methods, the proposed method can prescreen the database solely based on structural similarity. The proposed method can handle RNAs with arbitrary secondary structures, and it can be easily incorporated into various search methods that use different statistical models. Furthermore, the proposed approach has a low computational cost, yet very effective for prescreening, as will be demonstrated in the paper.**

I. Introduction

Recent studies on various genomes have revealed that there exist a large number of noncoding RNAs (ncRNAs), which are RNA molecules that function without being translated into proteins [6]. Unlike mRNAs (messenger RNAs) that passively carry protein-coding information, the ncRNAs actively participate in diverse biological processes. Although examples such as the tRNAs (transfer RNAs) and rRNAs (ribosomal RNAs) have been known for a long time, systematic research on ncRNAs shows that the number of RNAs and the variety and extent of their roles are much larger than it was previously thought [6].

As the annotation of ncRNAs is still at an early stage, it is of great importance to develop computational tools that can be used for screening the genome to identify new RNAs. An effective way for finding new ncRNAs is to search for RNAs that look similar to already known RNAs. Given a set of related RNAs that belong to the same family, we may construct a statistical model that represents the RNA family, and use this model to find similar regions in a genome database. This is usually called a *similarity search* or a *homology search.*

For a successful RNA similarity search, we need a suitable statistical model that can effectively describe the main characteristics of ncRNAs. One distinguishing feature of many ncRNAs is the conservation of their secondary structures. As the secondary structure of a ncRNA plays a crucial role in carrying out its biological function, many RNA families have characteristic secondary structures that are commonly shared by their members [1]. For this reason, it is important to consider both sequence similarity as well as structural similarity when performing an RNA similarity search. In fact, using a scoring scheme that can reasonably combine sequence and structural similarities can considerably increase the discriminative power of the search [1], [9].

The secondary structure of an RNA can be described in terms of correlations between distant bases in its primary sequence [9]. Therefore, in order to represent ncRNA families and develop a scoring scheme that can combine contributions from sequence similarity and structural similarity, we need a statistical model that can describe these base correlations. Examples of such models are the CM[1] (covariance model) [1] and the profile-csHMM (profile context-sensitive HMM) [8], [11]. Unfortunately, the computational cost for using these models is often too high for scanning a large genome database. In order to solve this problem, a number of methods have been proposed to expedite the RNA similarity search [3], [4], [7], [10]. The main idea underlying these methods is to prescreen the database using a simpler model (e.g., a profile-HMM) to filter out the dissimilar regions as much as possible. Only the remaining regions that bear some similarity are passed to a more complex (hence, more discriminative) model, such as the profile-csHMM or the CM, for further inspection. It has been shown that this prescreening approach can make the search speed significantly faster, either without any loss of accuracy [3], [4], [10] or at a slight loss of accuracy [7].

In this paper, we propose a novel prescreening method based on matched filtering of stem patterns. Unlike the previous methods, the proposed method can scan the database solely based on structural similarity, which can be especially useful for RNA families with low sequence similarity. As the

[1]A CM can be viewed as a SCFG (stochastic context-free grammar) with a special structure.

978-1-4244-2110-7/08/$25.00 ©2007 IEEE

matched filter is constructed from the secondary structure of the reference RNA instead of a specific statistical model, the proposed method can be used in combination with any kind of model, including profile-csHMMs and CMs. Furthermore, as the matched filter can be constructed for any kind of RNA secondary structure, the proposed method can be used for searching any RNA family, including those with pseudoknots[2] Finally, the proposed method has a low computational cost, yet very effective in detecting the structural similarity, as will be demonstrated in our experiments.

This paper is organized as follows. In Sec. II, we present a brief review of the existing prescreening methods. The proposed prescreening method based on structural similarity is elaborated in Sec. III. We present experimental results in Sec. IV that demonstrate the effectiveness of the proposed method, and the paper is concluded in Sec. V.

II. Fast RNA Search Using Prescreening Filters

A typical RNA similarity search is carried out as follows. Given a set of related RNAs that belong to the same family, we first find their multiple sequence alignment based on their sequence and/or structural similarity. There exist many heuristic methods that can be used to find a reasonably good alignment of the given sequences [2]. Based on this multiple sequence alignment, we predict the common secondary structure of the RNAs and construct a statistical model – such as a profile-csHMM [8], [11] or a CM (covariance model) [1] – that can closely represent the alignment.[3]

Once the model is constructed, it can be used for searching a genome database to find similar sequences which might be new members of the same RNA family. Given a target RNA, we compute a similarity score based on the constructed model, to find out how much it resembles the reference RNA family. The observation probability of the target RNA is a common choice for the similarity score, although it is typical to use either the log-probability or a log-likelihood ratio after normalizing the log-probability with respect to a random sequence model. Unfortunately, computing the observation probability of a target RNA based on profile-csHMMs or CMs is computationally expensive. This is due to the complexity of these models that is necessary for describing the complicated base correlations in RNA sequences. For example, the computational complexity of the optimal alignment algorithm[4] for CMs – called the *CYK (Cocke-Younger Kasami) algorithm* – is $O(L^3M)$, where L is the length of the target RNA and M is the number of states in the model. The number of states

[2]RNA secondary structures that have crossing base-pairs are called pseudoknots.

[3]Although we have described the procedure in three separate steps (i.e., finding the alignment, predicting the common secondary structure, and constructing the model) for simplicity, these steps are closely interrelated and it is typical to repeat these steps until the model converges to the optimal one.

[4]In general, there can be many different state sequences (or "paths") that give rise to the same symbol sequence. An *optimal alignment* algorithm tries to find the optimal path among all feasible paths that maximizes the observation probability of the sequence based on the given model. As this is conceptually identical to finding the best alignment between the symbol sequence and the statistical model, it is typically called an optimal alignment algorithm.

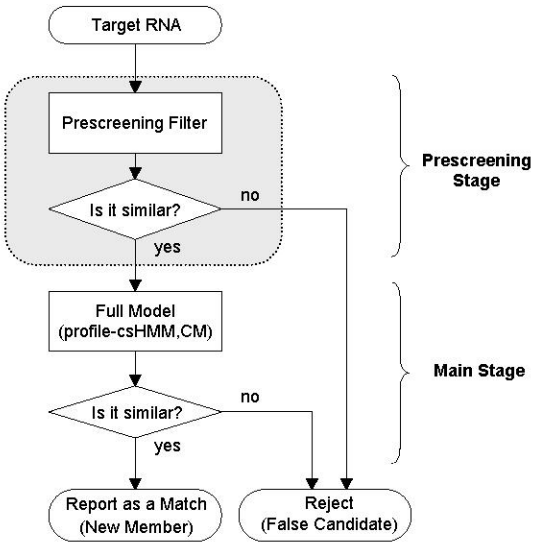

Fig. 1. Using a prescreening method can make the search significantly faster.

M is proportional to the length of the reference RNA. Similarly, the *SCA (sequential component adjoining) algorithm* for profile-csHMMs has a high computational complexity. The complexity of the SCA algorithm is variable, typically between $O(L^2M)$ and $O(L^4M)$, depending on the structure of the reference RNA. Compared to $O(LM)$ of the Viterbi algorithm[5] for profile-HMMs, the computational complexity of the CYK and SCA algorithms is relatively high. Because of the high computational cost of these algorithms, RNA search based on profile-csHMMs and CMs are usually too slow for scanning a large database, especially if the size of the RNA is large.

In order to overcome this problem, a general prescreening method has been proposed in [3] to make CM-based searches faster. The basic idea of the prescreening method is as follows. Instead of using a CM to scan the entire database, the method first uses a simple "prescreening filter" to scan the database.[6] The prescreening filter quickly filters out regions that are dissimilar, and passes only the regions that are similar enough to the reference RNA family to the second stage. In the second stage, a full CM is used to investigate these regions more closely. The basic idea of the prescreening approach is illustrated in Fig. 1. If the prescreening filter runs fast enough compared to the more complex model (a CM, in this case) yet effective enough to filter out most of the dissimilar regions, the overall speed of the search can be improved significantly. In [3], profile-HMMs were used as prescreening filters, whose parameters were chosen based on the CM parameters such that it guarantees that there is no loss in the prediction accuracy. It was shown that the search speed could be improved by 25 times on average, and by more than 200

[5]The complexity of the Viterbi algorithm for general HMMs is $O(LM^2)$.

[6]This should not be confused with the filters in signal processing. The prescreening filters are in fact simple statistical models that are used to "filter out" the dissimilar regions, hence called filters.

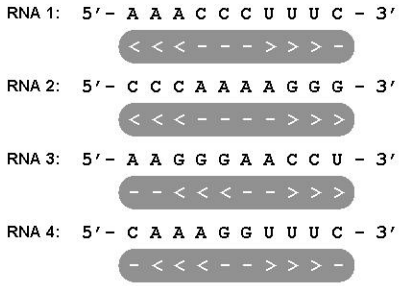

Fig. 2. Examples of RNAs with similar secondary structures.

times for many ncRNA familes [4]. Similarly, a prescreening method for profile-csHMMs was proposed in [10], which also used profile-HMMs for prescreening. Unlike CMs that cannot be used for finding RNAs with pseudoknots, profile-csHMMs have the advantage that they can represent any kind of RNAs, including pseudoknots. The original prescreening method proposed in [3] was improved further. For example, in [4], the profile-HMM based filters were augmented with some secondary structure information (though limited), and in [7], profile-HMM based heuristic filters were proposed that allow us to trade prediction accuracy for speed.

One disadvantage of the previous methods is that they mainly rely on sequence similarity. Although the method proposed in [4] augments the profile-HMM with sub-CMs (parts of the full CM) to detect simple stem-loops (hairpins), this hybrid prescreening filter can represent the structure of the reference RNA only partially. Furthermore, although the hybrid filter will still be faster than the full CM, it is considerably slower than the one solely based on the profile-HMM.

In the following section, we propose a novel prescreening method that effectively overcomes the shortcomings of the previous methods. The proposed method is based on matched filtering of stem patterns, which can scan the database to find regions that are structurally similar to the reference RNA. The structural matched filter can be constructed for any kind of RNA secondary structure, making the proposed approach generally applicable. Furthermore, it has a very low computational cost, hence suitable for scanning large databases.

III. Matched Filtering for Structural Similarity

Assume that we have a reference RNA with a known secondary structure. Given a target RNA sequence with no structural annotation, how can we quickly find out if a similar structure can be also found in the target RNA? For example, let us consider the RNAs shown in Fig. 2. As we can see in Fig. 2, all four RNAs have similar secondary structures, where every RNA has a single stem-loop at a different location and/or a different loop size. Their structural similarity can be easily recognized if we draw the dot-plots for their structures. Given an RNA sequence $\mathbf{x} = x_1 x_2 \dots x_L$ with a structural annotation, we define the corresponding $L \times L$ dot-plot matrix $\mathbf{P}$ as follows. The (m,n)-th element p_{mn} of the dot-plot

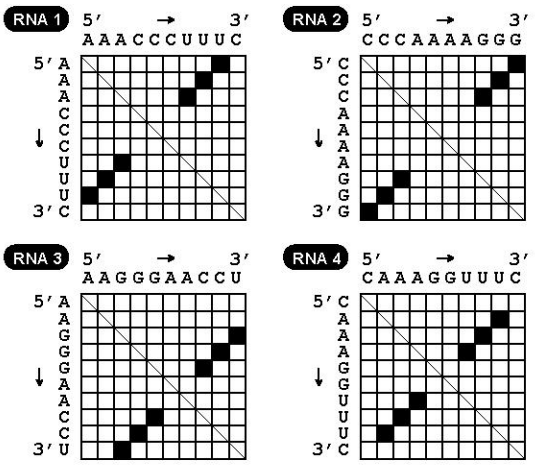

Fig. 3. The dot-plots corresponding to the RNAs in Fig. 2.

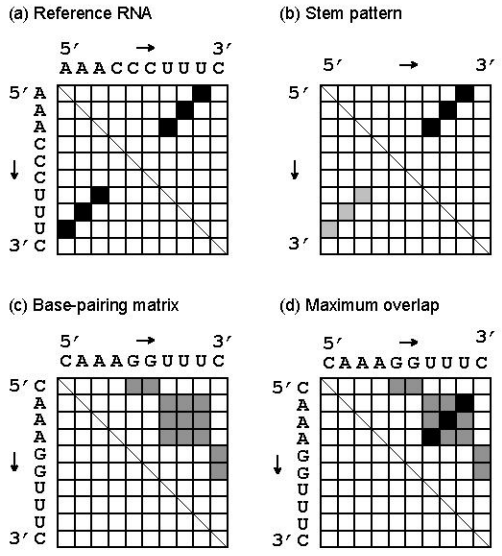

Fig. 4. Detecting structural similarity by comparing the dot-plot patterns.

matrix (base-pairing matrix) is defined as

$$p_{mn} = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ form a base-pair} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The dot-plot matrix $\mathbf{P}$ is always symmetric by definition. Fig. 3 shows the dot-plots that correspond to the structure of the RNAs in Fig. 2. From Fig. 3, we can readily recognize the structural similarity among the four RNAs.

This shows that using dot-plots can be very useful in detecting the structural similarity between RNAs. To demonstrate this idea, let us consider the following example. Assume that RNA-1 in Fig. 2 is used as the reference, and we assume that we know its structure. Based on its secondary structure, let us construct the base-pairing matrix $\mathbf{P}_r$ as shown in Fig. 4(a). As the matrix $\mathbf{P}_r$ is symmetric, we keep only the upper-triangular portion of $\mathbf{P}_r$ to obtain $\bar{\mathbf{P}}_r$. This (strictly) upper-triangular matrix $\bar{\mathbf{P}}_r$ is shown in Fig. 4(b), where the removed

portion is shown in gray. Note that this matrix contains the structural pattern (or the stem pattern) of the reference RNA. Now assume that RNA-4 is the target RNA whose structure we do not know. In order to find out whether the target (RNA-4) has a similar structure as the reference (RNA-1), we compute the base-pairing matrix $\mathbf{P}_t$ of RNA-4. Since we do not know its structure, we cannot construct $\mathbf{P}_t = \{p_{mn}\}$ based on the actual base-pairing information. Instead, we set $p_{mn} = 1$ for all (m, n) where x_m *can* form a base-pair with x_n. For example, if $x_m = A$, then we let $p_{mn} = 1$ for every n that satisfies $x_n = U$. As $\mathbf{P}_t$ is also symmetric, we keep only the upper-triangular portion of $\mathbf{P}_t$, and denote it as $\bar{\mathbf{P}}_t$. The matrix $\bar{\mathbf{P}}_t$ is shown in Fig. 4(c). Now that we have $\bar{\mathbf{P}}_r$ and $\bar{\mathbf{P}}_t$, we can compare these matrices to find out whether the target RNA has a similar structure as the reference RNA. One way to do this is to find the *maximum overlap* between the matrices as illustrated in Fig. 4(d). As expected, the stem pattern in $\bar{\mathbf{P}}_r$ completely overlaps with the base-pairing region in $\bar{\mathbf{P}}_t$, showing that the target (RNA-4) has a (nearly) identical structure as the reference (RNA-1).

Based on this idea, we propose an efficient method for comparing the structural similarity between a structured reference RNA and an unstructured target RNA. The proposed method is as follows. Firstly, we construct a $L_r \times L_r$ base-pairing matrix $\mathbf{P}_r = \{p_{mn}\}$ based on the secondary structure of the reference RNA, where L_r is the length of the RNA. We keep the upper-triangular portion of $\mathbf{P}_r$ to construct an upper-triangular matrix $\bar{\mathbf{P}}_r = \{\bar{p}_{mn}\}$ as follows

$$\bar{p}_{mn} = \begin{cases} p_{mn}, & \text{if } m < n \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

From $\bar{\mathbf{P}}_r$, we construct the matched filter matrix $\mathbf{S} = \{s_{mn}\}$ such that $s_{mn} = \bar{p}_{(L_r-m)(L_r-n)}$. This is identical to keeping the lower-tringular portion of $\mathbf{P}_r$ to obtain $\mathbf{S}$. Secondly, for a target RNA of length L_t, we construct a $L_t \times L_t$ base-pairing matrix $\mathbf{P}_t$ for *all* possible base-pairs. As before, we take the upper-triangular portion of $\mathbf{P}_t$ to get $\bar{\mathbf{P}}_t$. Thirdly, in order to compare the structures of the RNAs, we find the maximum overlap between the matrix $\bar{\mathbf{P}}_r$, which contains the stem pattern of the reference RNA, and the matrix $\bar{\mathbf{P}}_t$, which shows the base-pairing region of the target RNA. The maximum overlap can be easily found by computing

$$\mathbf{Y} = \bar{\mathbf{P}}_t * \mathbf{S}, \tag{3}$$

and finding the largest element of $\mathbf{Y}$, where $\mathbf{A} * \mathbf{B}$ denotes the two-dimensional convolution of $\mathbf{A}$ and $\mathbf{B}$. For $\mathbf{Y} = \{y_{mn}\}$, we define λ to be the value of the largest element

$$\lambda = \max_{m,n} \left(y_{mn}\right). \tag{4}$$

This λ gives us the *maximum number of base-pairs* that the two RNAs have in common. The larger the value of λ, the closer will be the structure of the reference RNA and that of the target RNA. The process described so far can be viewed as "matched filtering" of a noisy signal (structural pattern of the target RNA) based on the shape of the original signal (structural pattern of the reference RNA). The entire matched filtering process is illustrated in Fig. 5.

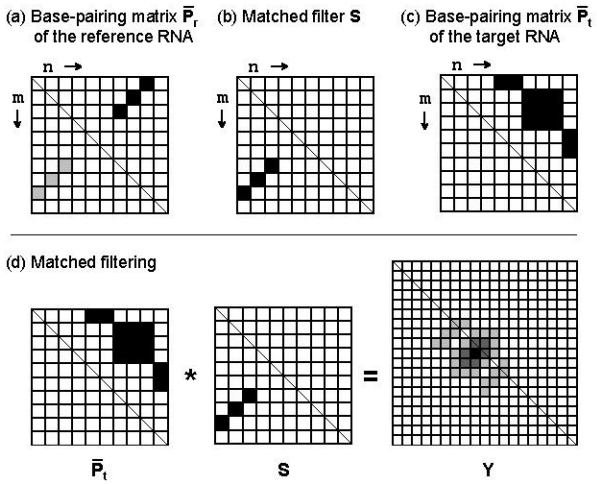

Fig. 5. Matched filtering based on stem patterns.

At first sight, the computational complexity of the proposed method seems to be $O(L_r^2 L_t^2)$ as it involves the convolution of an $L_r \times L_r$ matrix and an $L_t \times L_t$ matrix. However, the actual complexity is much smaller, because the matrix $\mathbf{S}$ is very sparse. In fact, the number of non-zero elements in $\mathbf{S}$ is identical to the number of base-pairs in the reference RNA. The actual computational complexity will be $O(NL_t^2)$, where N $(\leq \frac{1}{2}L_r)$ is the number of base-pairs. Although the complexity is still quadratic in L_t (length of the target RNA), this is not a serious problem in practice, as it simply corresponds to the addition of N matrices of size $L_t \times L_t$.

IV. Experimental Results

To demonstrate the effectiveness of the proposed method, we performed numerical experiments using the RNAs in the CORONA_PK3 and FLAVI_PK3 families in the Rfam database [5]. Note that both RNA families contain pseudo-knots, which cannot be represented by CMs though they can be represented by profile-csHMMs. The computational complexity of the SCA algorithm would be $O(L^4 M)$ for these RNA families [11]. Due to the high computational cost, an RNA search based on profile-csHMM alone would be too slow for scanning a large database, and it necessitates the incorporation of an efficient search strategy such as the prescreening approach.

In our experiments, we used the RNAs in the *seed alignments* [5] of the RNA families. For each family, we carried out the following cross-validation experiment. We first chose one of the members as the reference RNA, and constructed the matched filter matrix $\mathbf{S}$ based on its secondary structure. Using this matrix $\mathbf{S}$, we carried out the matched filtering process elaborated in Sec. III for the remaining members and computed λ as in (4). This value has been normalized to obtain the *normalized structural similarity score*

$$\sigma = \lambda / N, \tag{5}$$

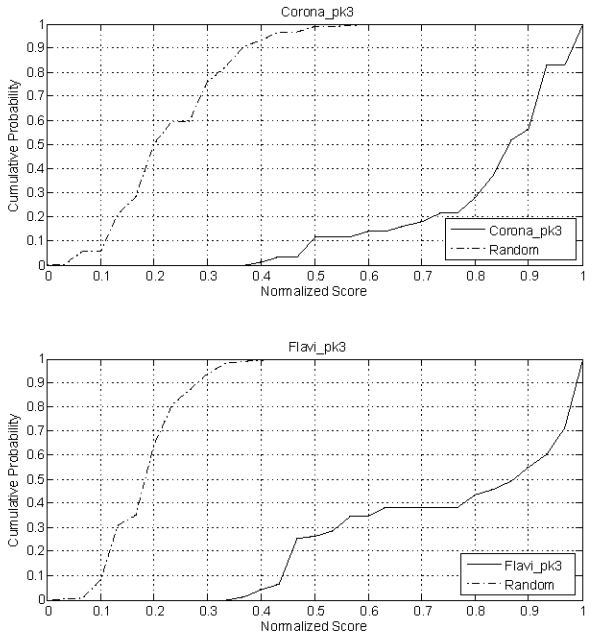

Fig. 6. Cumulative distribution of the structural similarity score. (Top) Score distribution of the CORONA_PK3 RNA family. (Bottom) Score distribution of the FLAVI_PK3 RNA family.

where N is the number of base-pairs in the reference RNA. Note that we always have $0 \leq \lambda \leq N$, hence the normalized score σ is in the region $0 \leq \sigma \leq 1$. This experiment has been repeated by using every member in the given family as the reference RNA, so that we can obtain a better estimate of σ. For comparison, we also computed the average structural similarity score for randomly generated RNA sequences.

If the score distribution of real RNAs (with similar secondary structures) is well-separated from that of random RNAs, we can use this score σ for filtering out the sequences that are structurally dissimilar from the reference RNA. In order to make this more reliable, we considered only stems with more than three base-pairs. Furthermore, we limited the region for finding the largest element of $\mathbf{Y}$ as follows

$$\lambda = \max_{|m-m_e|\leq D,\, |n-n_e|\leq D} \left(y_{mn}\right). \tag{6}$$

In (6), (m_e, n_e) is the expected location of the largest element for the case when the target RNA has an identical structure as the reference RNA. The parameter D restricts the region for finding the largest element around (m_e, n_e).

The experimental results are shown in Fig. 6, which shows the cumulative distribution function (CDF)

$$F_\sigma(s) = P(\sigma \leq s) \tag{7}$$

of the structural similarity score σ. Fig. 6 (Top) shows the score distribution of real RNAs and that of random RNAs, where the reference RNA family was the CORONA_PK3. As we can see in Fig. 6 (Top), the score distributions of real and random RNAs are well-separated.[7] For example, if we choose a threshold value of $\sigma^* = 0.45$, prescreening based on the proposed method can filter out 97% of the unrelated RNAs at a false negative prediction rate of 3% (i.e., 97% sensitivity). Considering that the proposed method performs much faster than the SCA algorithm for finding the optimal alignment, a rejection rate of 97% leads to a $33\,(= 1/0.03)$ times increase in the average search speed. For FLAVI_PK3 family, the proposed method performed even better. From Fig. 6 (Bottom) we can see that if we choose the threshold as $\sigma^* = 0.33$, more than 98% of the unrelated RNAs can be rejected at no loss of sensitivity. In this case, the search speed can be made around 50 times faster without any loss in the prediction performance.

[7] We used $D = 3$ in our experiments.

V. CONCLUDING REMARKS

In this paper, we proposed an efficient method for comparing the structural similarity of RNAs, based on matched filtering of stem patterns. The proposed method has a very low computational cost, yet it is applicable to RNAs with arbitrary secondary structures. As demonstrated in this paper, the proposed method can make RNA searches faster by prescreening the database based on structural similarity. Experimental results show that the search speed can be improved up to 50 times by the proposed method alone. We expect that the search speed can be improved even further when combined with other sequence-based prescreening methods.

ACKNOWLEDGMENT

This work was supported in part by the NSF grant CCF-0636799.

REFERENCES

[1] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge Univ. Press, Cambridge, UK, 1998.

[2] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment", *Current Opinion in Structural Biology*, vol. 16, pp. 368-373, 2006.

[3] Z. Weinberg and W. L. Ruzzo, "Faster genome annotation of non-coding RNA families without loss of accuracy", *Proc. 8th Ann. Int. Conf. on Computational Molecular Biology (RECOMB)*, pp. 243-251, 2004.

[4] Z. Weinberg and W. L. Ruzzo, "Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy", *Bioinformatics*, vol. 20, pp. i334-i340, 2004.

[5] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes", *Nucleic Acids Research*, vol. 33, pp. D121-D124, 2005.

[6] G. Storz, "An expanding universe of noncoding RNAs", *Science*, vol. 296, pp. 1260-1263, 2002.

[7] Z. Weinberg and W. L. Ruzzo, "Sequence-based heuristics for faster annotation of noncodingRNA families," *Bioinformatics*, vol. 22, pp. 35-39, 2006.

[8] B.-J. Yoon and P. P. Vaidyanathan, "Profile context-sensitive HMMs for probabilistic modeling of sequences with complex correlations", *Proc. 31st ICASSP*, Toulouse, May 2006.

[9] B.-J. Yoon and P. P. Vaidyanathan, "Computational identification and analysis of noncoding RNAs - Unearthing the buried treasures in the genome", *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 64-74, Jan. 2007.

[10] B.-J. Yoon and P. P. Vaidyanathan, "Fast search of sequences with complex symbol correlations using profile context-sensitive HMMs and pre-screening filters", *Proc. 32nd ICASSP*, Honolulu, Hawaii, Apr. 2007.

[11] B.-J. Yoon and P. P. Vaidyanathan, "Structural alignment of RNAs using profile-csHMMs and its application to RNA homology Search: Overview and new results", *IEEE Trans. Automatic Control & IEEE Trans. Circuits and Systems: Part-I (Joint Special Issue on Systems Biology)*, to appear in Jan. 2008.